

Detecting non-sinusoidal periodicities in observational data using multi-harmonic periodograms

Roman V. Baluev^{*}

Sobolev Astronomical Institute, St Petersburg State University, Universitetskij prospekt 28, Petrodvorets, St Petersburg 198504, Russia

Accepted 2009 February 12. Received 2009 February 9; in original form 2008 November 5

ABSTRACT

We address the problem of assessing the statistical significance of candidate periodicities found using the so-called ‘multi-harmonic’ periodogram, which is being used for detection of non-sinusoidal signals, and is based on the least-squares fitting of truncated Fourier series. The recent investigation (Baluev 2008) made for the Lomb-Scargle periodogram is extended to the more general multi-harmonic periodogram. As a result, closed and efficient analytic approximations to the false alarm probability, associated with multi-harmonic periodogram peaks, are obtained. The resulting analytic approximations are tested under various conditions using Monte Carlo simulations. The simulations showed a nice precision and robustness of these approximations.

Key words: methods: data analysis - methods: statistical - surveys

1 INTRODUCTION

The Lomb (1976)-Scargle (1982) (hereafter LS) periodogram is a well-known powerful tool, which is widely used to search for periodicities in observational data. The main idea used in the LS periodogram is to perform a least-squares fit of the data with a sinusoidal model of the signal and then to check how much the resulting weighted r.m.s. have decreased for a given signal frequency. The maximum value of the LS periodogram (i.e., the maximum decrement in the least-squares goodness-of-fit measure) corresponds to the most likely frequency of the periodic signal. This natural idea is quite easy to implement in numerical calculations.

However, random errors in the input data inspire noise peaks on the periodogram, so that we can never be completely sure that the peak that we actually observed was produced by a real periodicity. The common way to assess the statistical significance of the observed peak is based on the associated ‘false alarm probability’ (hereafter FAP). The FAP is the probability that the observed or larger periodogram peak could be produced by random measurement errors. The smaller is FAP, the larger is the statistical significance. Given some tolerance value FAP_* (say, 1%), we could claim that the detected candidate periodicity is statistically significant (if $FAP < FAP_*$) or is not (if $FAP > FAP_*$).

From the statistical viewpoint, the FAP is tightly connected with the probability distribution of periodogram maxima, which are calculated within some *a priori* fixed frequency segment.¹ However, even approximate calculation of

this distribution is a non-trivial task. It represented a trouble for astronomers for about three decades. It is worthwhile to mention here, for instance, the papers by Horne & Baliunas (1986); Koen (1990); Schwarzenberg-Czerny (1998a,b); Cumming et al. (1999); Cumming (2004); Frescura et al. (2008). Recently, a significant progress in this field was attained in the paper (Baluev 2008), where closed and simultaneously rather efficient approximations of the FAP for the LS periodogram are given, basing on results in the theory of extreme values of stochastic processes.

However, periodic signals being dealt with in astronomy often are significantly non-sinusoidal. Then the use of the LS periodogram is not optimal, since the corresponding periodic variation would be fitted inadequately. For instance, it is the case for lightcurves of variable stars of several types and for radial velocity curves of stars orbited by a planet on an eccentric orbit. Several ways to deal with this issue were proposed (for further references see e.g. Schwarzenberg-Czerny 1998a,b). In this paper, we focus attention on the so-called multi-harmonic periodogram (Schwarzenberg-Czerny 1996), which is based on the least-squares fitting of truncated Fourier series. Note that in the paper (Baluev 2008) a general class of periodograms based on the least-squares data fitting was considered as well, but from theoretical positions only. Here our aim is to apply these general results to the multi-harmonic periodograms.

The plan of the paper is as follows. In Section 2, we

^{*} E-mail: roman@astro.spbu.ru

¹ Speaking more precisely, the periodogram maxima are always

formulate the problem rigorously and introduce the necessary mathematical definitions. In Section 3, basing on the work (Baluev 2008), we derive closed approximations of the FAP, associated with multi-harmonic periodogram peaks. In Section 4, we use numerical Monte Carlo simulations to quantify the accuracy of these analytic approximations.

2 GENERAL DEFINITIONS

Let us write down the temporal model of the putative periodic signal using a trigonometric polynomial of some *a priori* stated degree n :

$$\mu(t, \boldsymbol{\theta}, f) = \sum_{k=1}^n (a_k \cos 2\pi k f t + b_k \sin 2\pi k f t), \quad (1)$$

where f is the signal frequency and the vector $\boldsymbol{\theta}$ incorporates $d = 2n$ Fourier coefficients a_k, b_k . Further we adopt exactly the same notations as those used in (Baluev 2008). Clearly, the model μ is linear: $\mu(t, \boldsymbol{\theta}, f) = \boldsymbol{\theta} \cdot \boldsymbol{\varphi}(t, f)$, where the vector $\boldsymbol{\varphi}(t, f)$ incorporate the first n harmonics of the Fourier basis. In addition to the signal model μ , we define the base temporal model $\mu_{\mathcal{H}}(t, \boldsymbol{\theta}_{\mathcal{H}}) = \boldsymbol{\theta}_{\mathcal{H}} \cdot \boldsymbol{\varphi}_{\mathcal{H}}(t)$, which is assumed to be linear with respect to $d_{\mathcal{H}}$ unknown parameters $\boldsymbol{\theta}_{\mathcal{H}}$. This base model may represent, for instance, a constant or a long-term polynomial (e.g., linear or quadratic) temporal trend. Therefore, the alternative (full) model is given by $\mu_{\mathcal{K}}(t, \boldsymbol{\theta}_{\mathcal{K}}, f) = \mu_{\mathcal{H}}(t, \boldsymbol{\theta}_{\mathcal{H}}) + \mu(t, \boldsymbol{\theta}, f)$, where $\boldsymbol{\theta}_{\mathcal{K}}$ incorporates all parameters in $\boldsymbol{\theta}_{\mathcal{H}}$ and $\boldsymbol{\theta}$. From the viewpoint of the statistical tests theory, we need to test the base hypothesis $\mathcal{H} : \boldsymbol{\theta} = 0$ against the alternative one $\mathcal{K} : \boldsymbol{\theta} \neq 0$.

The input dataset consists of N measurements x_i taken at timings t_i and having uncertainties σ_i . We assume that the random errors of the measurements are statistically independent and normally distributed. Below we will deal with the least-squares periodograms defined in (Baluev 2008). These periodograms are based on the linear least-squares fitting procedure. The basic one, $z(f)$, represents the half-difference

$$z(f) = [\chi_{\mathcal{H}}^2 - \chi_{\mathcal{K}}^2(f)] / 2, \quad (2)$$

where $\chi_{\mathcal{H}}^2$ and $\chi_{\mathcal{K}}^2$ represent the minimum values of the χ^2 goodness-of-fit statistic, calculated under the two corresponding hypotheses, \mathcal{H} and \mathcal{K} . Note that under the base hypothesis \mathcal{H} the random quantities $\chi_{\mathcal{H}}^2$ and $\chi_{\mathcal{K}}^2$ follow the χ^2 -distributions with $N_{\mathcal{H}} = N - d_{\mathcal{H}}$ and $N_{\mathcal{K}} = N - d_{\mathcal{K}}$ degrees of freedom and thus indeed represent χ^2 -variates. The periodogram $z(f)$ can be only calculated if the variances σ_i of the observational errors are known exactly. Usually we do not know these variances exactly, and can fix only the statistical weights, $w_i \propto 1/\sigma_i^2$, so that $\sigma_i^2 = \kappa/w_i$ with the common factor κ being unconstrained *a priori*. Therefore, we will also consider three modified least-squares periodograms:

$$\begin{aligned} z_1(f) &= N_{\mathcal{H}} \frac{\chi_{\mathcal{H}}^2 - \chi_{\mathcal{K}}^2(f)}{2\chi_{\mathcal{H}}^2}, \quad z_2(f) = N_{\mathcal{K}} \frac{\chi_{\mathcal{H}}^2 - \chi_{\mathcal{K}}^2(f)}{2\chi_{\mathcal{K}}^2(f)}, \\ z_3(f) &= \frac{N_{\mathcal{K}}}{2} \ln \frac{\chi_{\mathcal{H}}^2}{\chi_{\mathcal{K}}^2(f)}. \end{aligned} \quad (3)$$

These periodograms do not depend on κ and can be calculated even if κ is unknown. The periodograms $z_1(f)$ and $z_2(f)$ represent normalizations of the basic periodogram

$z(f)$ by the sample variances of the residuals, calculated under one of the two hypotheses, \mathcal{H} or \mathcal{K} . The periodogram $z_3(f)$ is proportional to the logarithm of the likelihood ratio statistic. More discussion of these definitions can be found in (Baluev 2008). A discussion of several issues associated with the least-squares interpretation of the periodograms introduced above can be also found in (Schwarzenberg-Czerny 1998a,b; Zechmeister & Kürster 2009). The modified periodograms $z_{1,2,3}$ are unique-value monotonic functions of each other and thus are entirely equivalent for the practical use.

The definitions (3) imply that the statistical weights w_i should be known with sufficient precision, and only the proportionality factor is unknown. This framework is a usually adopted for the period analysis of astronomical data (e.g. Gilliland & Baliunas 1987; Irwin et al. 1989; Zechmeister & Kürster 2009) and we adopt it here. Nevertheless, sometimes this model may not work well. For instance, the paper (Baluev 2009) discusses the case in which the weights of observations are not known *a priori* with sufficient precision. In this case, the traditional multi-harmonic periodograms being discussed here may not work well.

We do not discuss in detail the numerical algorithms for calculation of the periodograms introduced. The form of the above definitions is more suitable for quantifying the statistical distributions of the corresponding periodograms. Fast numerical algorithms of practical evaluation of the multi-harmonic periodograms are given in (Schwarzenberg-Czerny 1996; Palmer 2009).

3 FALSE ALARM PROBABILITY

Let us pick any of the periodograms introduced above, and denote it as $Z(f)$. If the frequency of the putative signal was known, the false alarm probability $\text{FAP}_{\text{single}}(Z)$, associated with the given value $Z(f)$, could be calculated as $\text{FAP}_{\text{single}}(Z) = 1 - P_{\text{single}}(Z)$, where $P_{\text{single}}(Z)$ is the cumulative distribution of the corresponding periodogram value, calculated under the base hypothesis \mathcal{H} . It is well-known that within simple constant scale factors these distributions are $\chi^2(d)$, $F(d, N_{\mathcal{K}})$, and $B(d, N_{\mathcal{K}})$ for the periodograms z , z_2 , and z_1 , respectively (see, e.g., Schwarzenberg-Czerny 1998a,b; Baluev 2008). Here the quantities in brackets mark the necessary numbers of degrees of freedom.

When the signal frequency is unknown *a priori*, we need to search for a maximum of $Z(f)$ within some wide frequency band $[f_{\min}, f_{\max}]$. From now on we will assume, for the sake of definiteness, that $f_{\min} = 0$. In practice it is a frequent case and also this assumption allows us to simplify the formal expressions. All results presented below can be easily extended to the case of arbitrary $f_{\min} > 0$. For example, we will need to replace certain lower integration limits appropriately and to change the expressions for the frequency bandwidth from f_{\max} to $f_{\max} - f_{\min}$. According to Baluev (2008), to estimate the FAP associated with the observed maximum, we use the Davies (1977, 1987, 2002) bound

$$\text{FAP}_{\max}(Z, f_{\max}) \leq \text{FAP}_{\text{single}}(Z) + \tau(Z, f_{\max}), \quad (4)$$

Exact expressions for function τ are given in (Baluev 2008) for the general least-squares periodogram z and for its modifications $z_{1,2,3}$ (see eqs. (7) and (8) in that paper). In fact,

the right hand side in the inequality (4) represents something more than just an upper bound. It was demonstrated by Baluev (2008), that in the LS periodogram case the inequality (4) appears rather sharp, especially for practically important low FAP levels. In addition to the bound (4), we will deal with the following approximation:

$$\begin{aligned} \text{FAP}_{\max}(Z, f_{\max}) &= 1 - P_{\max}(Z, f_{\max}), \\ P_{\max}(Z, f_{\max}) &\approx e^{-\tau(Z, f_{\max})} P_{\text{single}}(Z). \end{aligned} \quad (5)$$

As it was discussed in (Baluev 2008), the formulae (5) should provide a good approximation to FAP_{\max} *uniformly* (i.e., for all FAP levels) in the case of small aliasing. Note that the approximation (5) and the bound (4) yield almost coinciding results if $\text{FAP} < 0.1$, so that the mentioned property of the approximation (5) probably will not have direct practical application. In this paper, we use (5) just to plot a reference ‘alias-free’ FAP curve.

Now we need to obtain the function $\tau(z, f_{\max})$ for our special case of the multi-harmonic periodograms. In particular, we need to calculate the factor $A(f_{\max})$, present in the expressions for τ . In general, this factor depends in a rather unpleasant way on the models of the data, on the time series sampling, and on the sequence of the statistical weights of observations. To attain some technical simplicity, let us firstly assume that, like in the classical LS periodogram, the base model is empty: $d_{\mathcal{H}} = 0$ and $\mu_{\mathcal{H}}(t) \equiv 0$. In this case, we need to find firstly the eigenvalues λ_k of the $d \times d$ matrix \mathbf{M} , which is defined as:

$$\begin{aligned} \mathbf{Q} &= \overline{\varphi \otimes \varphi}, \quad \mathbf{S} = \overline{\varphi \otimes \varphi'_f}, \\ \mathbf{R} &= \overline{\varphi'_f \otimes \varphi'_f}, \quad \mathbf{M} = \mathbf{Q}^{-1}(\mathbf{R} - \mathbf{S}^T \mathbf{Q}^{-1} \mathbf{S}). \end{aligned} \quad (6)$$

Here the overline denotes the weighted averaging over the time series and the binary operation \otimes is the dyadic product of vectors ($\mathbf{x} \otimes \mathbf{y} = \mathbf{xy}^T$), see Appendix A in (Baluev 2008). The notation φ'_f stands for the partial derivative of the vectorial function $\varphi(t, f)$ over f . We use the expressions from the paper (Davies 1987) to calculate the factor $A(f_{\max})$. We need to combine eqs. (3.2, 3.3) and the unnumbered equation following after the eq. (3.4) from (Davies 1987) to obtain the formula (7) in the paper (Baluev 2008) with

$$A = \frac{\pi^{n-1}}{\Gamma(n + \frac{1}{2})} \int_0^{f_{\max}} df \int_0^\infty \left(1 - \frac{1}{\prod_{k=1}^{2n} \sqrt{1 + x\lambda_k(f)}} \right) \frac{dx}{x^{3/2}}. \quad (7)$$

We need to obtain some more simple, although possibly approximate, expression for the factor A . To do this, we firstly obtain a suitable approximation to the matrix \mathbf{M} and hence to its eigenvalues λ_k . After that, we can substitute the approximations for λ_k to (7), in order to derive the final approximation to $A(f_{\max})$. We give the associated details, as well as an assessment of the practical precision of the resulting approximation, in the Appendix A. Here we give only the final result, which seems to be sufficiently accurate in practice. The matrix \mathbf{M} can be approximated by the following diagonal block form:

$$\mathbf{M} \approx \pi T_{\text{eff}}^2 \begin{pmatrix} \mathbf{I}_2 & 0 & \dots & 0 \\ 0 & 2^2 \mathbf{I}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & n^2 \mathbf{I}_2 \end{pmatrix}, \quad (8)$$

where \mathbf{I}_2 is the 2×2 identity matrix, T_{eff} is the effective time-

Table 1. The constants α_n for a few values of n .

n	1	2	3	5	8	15
α_n	1	1.556	1.062	0.136	$9.921 \cdot 10^{-4}$	$1.037 \cdot 10^{-10}$

span ($T_{\text{eff}} = \sqrt{4\pi \mathbb{D}t}$, where $\mathbb{D}t$ is the weighted variance of timings t_i , see Baluev 2008). The approximate equality (8) implies that the $2n$ eigenvalues required are grouped into n pairs $\lambda_{2k-1} \approx \lambda_{2k} \approx \pi T_{\text{eff}}^2 k^2$, $k = 1, 2, \dots, n$. Finally,

$$A(f_{\max}) \approx 2\pi^{n+\frac{1}{2}} \alpha_n W, \quad (9)$$

where $W = f_{\max} T_{\text{eff}}$ and

$$\alpha_n = \frac{2^n}{(2n-1)!!} \sum_{k=1}^n \frac{(-1)^{n-k} k^{2n+1}}{(n+k)!(n-k)!}. \quad (10)$$

Here the quantity $(2n-1)!!$ represents the product of all odd integers from $(2n-1)$ downto 1. The numerical values of the constants α_n for a few values of n are given in Table 1.

Therefore, using eqs. (7, 8) from (Baluev 2008), we obtain for the basic multi-harmonic periodogram $z(f)$

$$\tau \approx W \alpha_n e^{-z} z^{n-\frac{1}{2}}, \quad (11)$$

and for the associated modified periodograms $z_{1,2,3}(f)$

$$\tau \approx W \alpha_n \frac{\Gamma(\frac{N_{\mathcal{H}}}{2})}{\Gamma(\frac{N_{\mathcal{K}}+1}{2})} \times \begin{cases} \left(\frac{2z_1}{N_{\mathcal{H}}}\right)^{n-\frac{1}{2}} \left(1 - \frac{2z_1}{N_{\mathcal{H}}}\right)^{\frac{N_{\mathcal{K}}-1}{2}}, \\ \left(\frac{2z_2}{N_{\mathcal{K}}}\right)^{n-\frac{1}{2}} \left(1 + \frac{2z_2}{N_{\mathcal{K}}}\right)^{-\frac{N_{\mathcal{H}}}{2}+1}, \\ \left(2 \sinh \frac{z_3}{N_{\mathcal{K}}}\right)^{n-\frac{1}{2}} e^{-z_3 \left(1 + \frac{2n-3}{2N_{\mathcal{K}}}\right)}. \end{cases} \quad (12)$$

4 NUMERICAL SIMULATIONS

Speaking in terms of the statistical tests theory, two kinds of mistakes can be made in the signal detection problem: the false alarm and the false non-detection. Our primary goal was to keep the false alarm probability at some *a priori* small levels $\text{FAP} < \text{FAP}_*$. This is guaranteed by the theoretical inequality (4). Now our goal is to characterize (given the condition of bounded FAP) the detection power, which is provided by the actual precision of the FAP estimation. We noted above that the right hand side in (4) is expected to provide some approximation to the FAP, not just an upper bound. However, the error of this approximation depends on conditions: in the case when distant periodogram values are weakly correlated, this approximation should be precise, and in the case when there exist pairs (or more complicated combinations) of strongly correlated distant periodogram values, this precision decreases (see Baluev 2008, Appendix B). In practice, the absence of strongly correlated peaks means that the periodograms are free from aliases.

Since we have been already prevented (at the given probability FAP_*) from false alarms by the upper character of the Davies bound (4), now we are more interested in precise approximation of detection thresholds (i.e., such critical values z_* that provide $\text{FAP}(z_*) = \text{FAP}_*$) rather than of the FAPs themselves, because it is the detection threshold z_* that determine the detection probability. This means that we should pay major attention to horizontal deviations between the simulated and theoretical FAP curves, rather than to vertical ones.

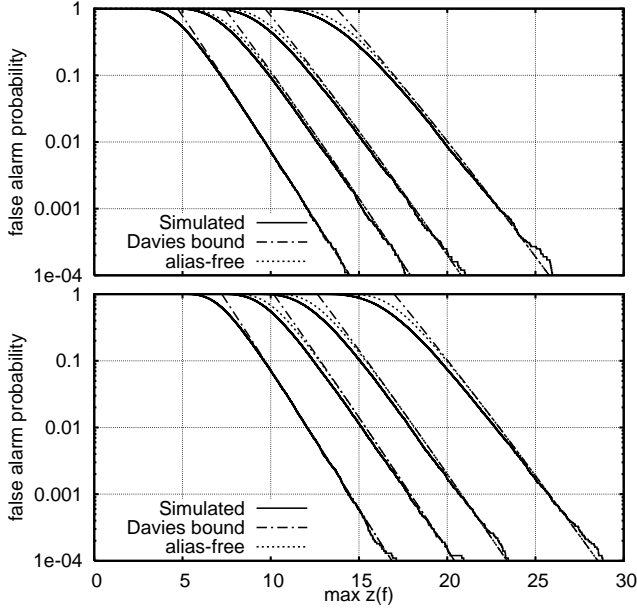


Figure 1. Simulated vs. analytic false alarm probability for the multi-harmonic periodogram of $N = 1000$ evenly spaced observations. Results for $n = 1, 2, 3, 5$ are shown as converging bunches of curves from left to right on each panel. The frequency bandwidth was $f_{\max}T = 50$ (top panel) and 500 (bottom panel). Here and in all other similar figures further, the number of Monte Carlo trials was about 10^5 for each simulation curve.

We now proceed to testing the precision of the theoretical approximations obtained above using Monte Carlo simulations of FAP_{\max} , in the same way as in (Baluev 2008). When the order of the approximating trigonometric polynomial grows, the volume of necessary calculations increases significantly due to the following reasons:

- (i) The calculations of single values of the multi-harmonic periodogram require to solve higher-dimensional linear least-squares problem (or to orthogonalize higher-dimensional functional bases).
- (ii) As the simulations have shown, the average density of peaks on the multi-harmonic periodograms increase roughly as $\mathcal{O}(n)$. This requires for the calculations to be performed on a more dense frequency grid, in order to obtain enough accurate values of periodogram maxima.

Therefore, our abilities in making numerical simulations are severely limited to small n only.

Firstly let us deal with the case when the time series does not produce any aliasing in the classical sense, i.e. on the LS periodogram. This is the case of a large number of evenly distributed observations. The corresponding simulated FAP_{\max} curves are shown in Fig. 1 for the periodograms $z(f)$. We can see that the theoretical approximations work quite well. Nevertheless, the small deviations for the cases $n \geq 2$ contrast with the LS case $n = 1$, for which we cannot see any deviation at all. Probably these small deviations emerged because of an extra correlation of distant periodogram values, caused by the fact that the model (1) incorporates several sinusoidal harmonics instead of one. Thus the periodogram values at two frequencies f_1 and f_2 appear correlated if $f_1/f_2 \approx p/q$ for some integers p, q not exceeding n (for $n = 1$ we had only the trivial condition

$f_1 \approx f_2$). Nevertheless, this subtle self-aliasing² effect seems to have negligible influence on the precision of our analytic FAP estimation, at least for $n \leq 5$. The corresponding errors of periodogram detection thresholds are about (or less) 2–3 per cent in these cases. Since the signal amplitude scales roughly as \sqrt{z} , this results in only ~ 1 per cent inaccuracy in the amplitude thresholds.

When the number of observations decreases and their temporal distribution becomes non-uniform, the precision of the analytic approximations for $n \geq 2$ decreases in the same manner as for the usual LS periodogram, $n = 1$. Fig. 2 shows a series of simulations for randomly spaced time series and for a time series with imposed periodic gapping of timings. We can see that the approximations of the threshold levels z_* , corresponding to $\text{FAP}_* \sim 0.01$, still are rather precise in many cases, which quite could correspond to a practical situation. The precision of the theoretical approximations decreases when f_{\max} or n grow, when N decreases, or when the degree of the non-uniformity of timings distribution increases. Nevertheless, even in the worst cases the relative error of z_* (corresponding to $\text{FAP}_* = 0.01$) does not exceed ~ 20 per cent, resulting in only ~ 10 per cent overestimation of the corresponding amplitude thresholds. Such loss of precision still is not catastrophic and quite can be tolerated.

The paper (Baluev 2008) in fact paid undeservedly small attention to the modified LS periodograms $z_{1,2,3}(f)$. It was assumed that the behaviour of their FAP curves is similar to the behaviour of the FAP curves of the basing LS periodogram. However, they are the modified periodograms which are usually used in practice. Here we try to correct this mistake. It appears that the FAP curves of modified periodograms are considerably less sensitive to an uneven time series sampling. Consequently, the precision of the Davies bound (4) and of the alias-free approximation (5) appears significantly better (see Fig. 3). For the modified multi-harmonic periodograms, the random distribution of timings does not introduce any significant perturbation of the FAP curve even for N as small as 30. In this case, the FAP curves for the modified periodograms perfectly agree with the alias-free approximation (5). For periodically gapped timings, the precision of the analytic FAP estimations improves too. Moreover, this precision does not decrease and even seem to *increase* when the order n or the frequency bandwidth f_{\max} grow. The reason for such refinement of the precision of the analytic estimations of the FAP for the modified periodograms is unclear.

It is harder to complete a similar series of Monte Carlo simulations for more complicated cases, e.g. with the base model $\mu_{\mathcal{H}}$ incorporating at least a constant or a linear trend. We present only a few examples of such simulations, which nevertheless further certify the practical efficiency of the closed expressions for the FAP described above (Fig. 4). Actually, it looks that a low-order polynomial trend in the base

² One may argue that such understanding of the notion ‘aliasing’ is not traditional, because the associated effect is not connected with uneven time series sampling, and is only a result of an interplay between the main period and its subharmonics. Nevertheless, for the sake of a uniform terminology, we name here all ‘wrong’ periodogram peaks as aliases, and the associated phenomenon of correlativity of distant periodogram values as aliasing.

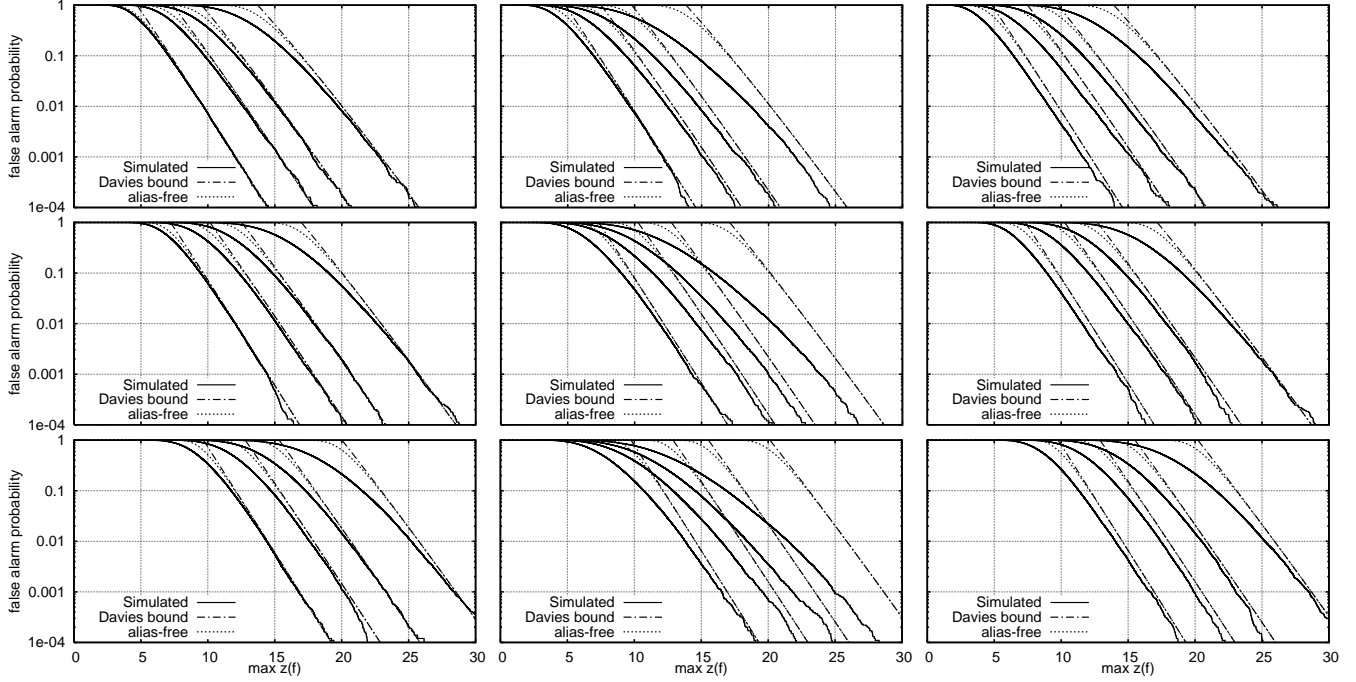


Figure 2. Simulated vs. analytic false alarm probability for the basic multi-harmonic periodogram $z(f)$. For the panels in the left and middle columns, the time series consisted of $N = 100$ and $N = 30$ randomly spaced datapoints, respectively. For the panels in the right column $N = 100$ datapoints were clumped in ten evenly spaced groups. Each group consisted of ten points and spanned only $1/50$ fraction of the total time-span (instead of the natural $1/10$ fraction). In each panel, four converging bunches of curves from left to right correspond to $n = 1, 2, 3, 5$. For the top row $f_{\max}T = 50$, for the middle one $f_{\max}T = 500$, and for the bottom one $f_{\max}T = 5000$.

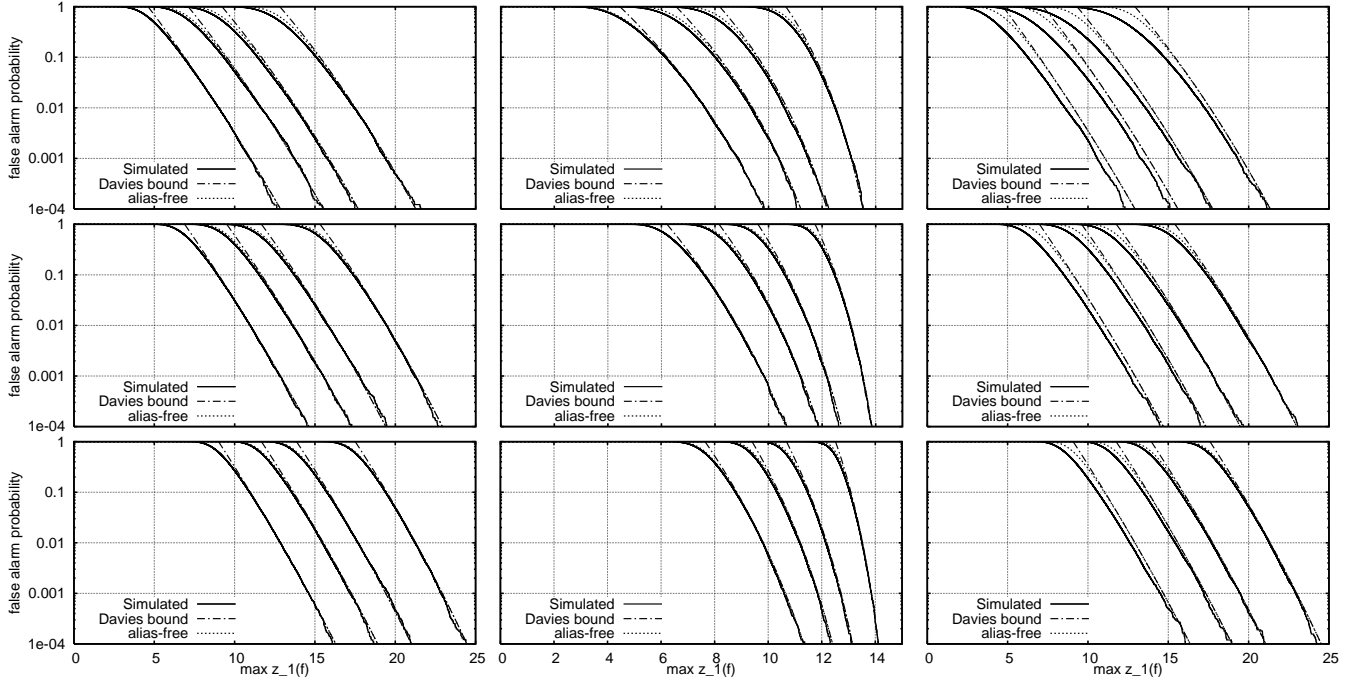


Figure 3. Same as in Fig. 2, but for the modified periodogram $z_1(f)$.

model $\mu_{\mathcal{H}}$ does not introduce any visible deviation in the simulated FAP curve, at least in this particular case.

Finally, let us take some realistic time series sampling and consider the associated FAP curves and their approximations under some realistic conditions. For this purpose, as

in the paper (Baluev 2008), we use the observational dates and standard errors of the high-precision radial velocity data for the stars 51 Peg and 70 Vir (Naef et al. 2004). The number of observations in the first time series is $N = 153$ and in the second one $N = 35$. The time-span of these time se-

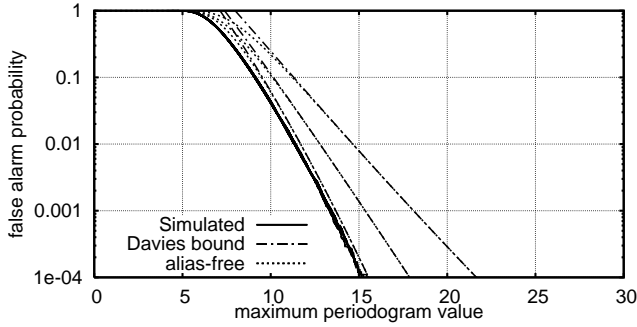


Figure 4. Simulated vs. analytic false alarm probability for the modified multi-harmonic ($n = 2$) periodogram $z_1(f)$, constructed from $N = 100$ evenly spaced observations, in the main frequency band $f_{\max}T = 50$. The graph shows *four* simulated FAP curves for different degrees of the polynomial trend in the base model $\mu_{\mathcal{H}}$: empty base model ($d_{\mathcal{H}} = 0$), a constant term ($d_{\mathcal{H}} = 1$), a linear trend ($d_{\mathcal{H}} = 2$), a quadratic trend ($d_{\mathcal{H}} = 3$). All these curves appear almost coinciding. For an intercomparison, we show here the theoretical distribution curves for all the modified periodograms, z_1 , z_3 , and z_2 (from left to right). Note that we plot them only for the case $d_{\mathcal{H}} = 0$, because the similar curves for $d_{\mathcal{H}} = 1, 2, 3$ did not show any visible deviation.

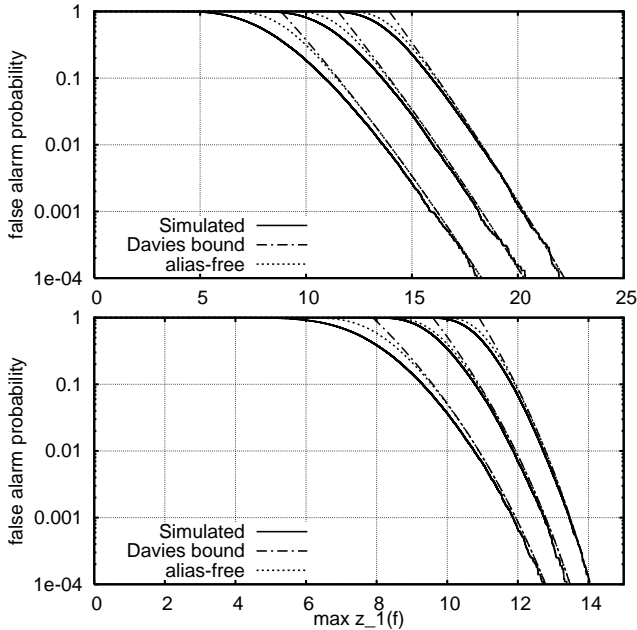


Figure 5. Simulated vs. analytic false alarm probability for the multi-harmonic ($n = 3$) periodogram $z_1(f)$ constructed from the radial velocity time series of 51 Peg (top) and 70 Vir (bottom). The base model $\mu_{\mathcal{H}}$ incorporated a free constant term ($d_{\mathcal{H}} = 1$). In each panel, three bunches of curves correspond to $P_{\min} = 1/f_{\max} = 100$ days, 10 days, and 1 day (from left to right).

ries are about a decade. In both cases, significant aliasing is present (e.g., corresponding to the annual and diurnal periods). We can see, however, that in both cases the analytic formulae for the periodogram z_1 work very well (Fig. 5).

It is worth noting that in all the cases discussed above, the Davies bound (4) indeed bounds the simulated FAP curves from the upper-right side, at least for not very small

levels $\text{FAP} > 10^{-3}$, which can be reliably modelled using 10^5 Monte Carlo trials.

5 CONCLUSIONS

In this paper, previous results by Baluev (2008) are applied to the case when the model of the signal to be detected represents a truncated Fourier polynomial. Closed analytic expressions for the false alarm probabilities, associated with multi-harmonic periodogram peaks, are given. They are tested under various conditions using Monte Carlo simulations. The simulations have shown that the accuracy of the mentioned theoretical estimations usually is quite suitable in practice. Also, these simulations have revealed an unexpected (but pleasant) phenomenon: the accuracy of the above theoretical approximations of the FAP is considerably better for the normalized multi-harmonic periodograms than for the basic, purely least-squares, ones. Since in practice the observational noise variance is rarely known precisely, they are the normalized periodograms that are usually dealt with. Therefore, the better behaviour of the FAP curves for the normalized periodograms has high practical value.

The necessary amount of Monte Carlo simulations for the multi-harmonic periodogram is bigger than for the LS one. It may appear very difficult to obtain a sufficiently precise Monte Carlo estimation of the false alarm probability even in the case of a single time series. Most likely, for surveys dealing with large numbers of separate time series, it would be impossible to perform the necessary amount of Monte Carlo simulations. For example, a single CPU at 2 GHz would complete all Monte Carlo simulations presented above in a few months only. On contrary, the closed theoretical estimations presented in this paper do not require any simulations at all, and simultaneously often have a nice accuracy. This indicates that the mentioned estimations represent a promising practical tool and may be used in a wide variety of astronomical applications, involving search for non-sinusoidal periodicities in observational data. The corresponding research fields are ranged from the studies of variable stars to the studies of extrasolar planetary systems.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (Grant 09-02-00230) and by the Russian President Programme for the State Support of Leading Scientific Schools (Grant NSh-1323.2008.2). I am grateful to the anonymous referee for providing important suggestions, which helped to improve the manuscript.

REFERENCES

- Baluev R. V., 2008, MNRAS, 385, 1279
- Baluev R. V., 2009, MNRAS,
doi:10.1111/j.1365-2966.2008.14217.x
- Cumming A., 2004, MNRAS, 354, 1165
- Cumming A., Marcy G. W., Butler R. P., 1999, ApJ, 526, 890
- Davies R. B., 1977, Biometrika, 64, 247
- Davies R. B., 1987, Biometrika, 74, 33

- Davies R. B., 2002, *Biometrika*, 89, 484
 Frescura F. A. M., Engelbrecht C. A., Frank B. S., 2008, *MNRAS*, 388, 1693
 Gilliland R. L., Baliunas S. L., 1987, *ApJ*, 314, 766
 Horne J. H., Baliunas S. L., 1986, *ApJ*, 302, 757
 Irwin A. W., Campbell B., Morbey C. L., Walker G. A. H., Yang S., 1989, *PASP*, 101, 147
 Koen C., 1990, *ApJ*, 348, 700
 Lomb N. R., 1976, *Ap&SS*, 39, 447
 Naef D., Mayor M., Beuzit J. L., Perrier C., Queloz D., Sivan J. P., Udry S., 2004, *A&A*, 414, 351
 Palmer D. M., 2009, *ApJ*, accepted, arXiv: 0901.1913
 Scargle J. D., 1982, *ApJ*, 263, 835
 Schwarzenberg-Czerny A., 1996, *ApJ*, 460, L107
 Schwarzenberg-Czerny A., 1998a, *MNRAS*, 301, 831
 Schwarzenberg-Czerny A., 1998b, *Baltic Astron.*, 7, 43
 Zechmeister M., Kürster M., 2009, *A&A*, accepted, arXiv: 0901.2573

APPENDIX A: THE FACTOR $A(F_{\text{MAX}})$

The elements of the matrices $\mathbf{Q}, \mathbf{S}, \mathbf{R}$ can be transformed in the way similar to eqs. (10) in (Baluev 2008). The matrix \mathbf{Q} will contain the averages of the kind $\overline{\sin k\omega t}$ and $\overline{\cos k\omega t}$. The matrix \mathbf{S} will contain components $\overline{t \sin k\omega t}$, $\overline{t \cos k\omega t}$, and also \overline{t} . The matrix \mathbf{R} will contain components of the kind $\overline{t^2 \sin k\omega t}$, $\overline{t^2 \cos k\omega t}$, and also $\overline{t^2}$. Here $k = 1, 2, \dots, 2n$. Therefore, we deal with quantities having the form

$$\begin{aligned} \Omega_s(f) &= \overline{\sin \omega t}, & \Omega_c(f) &= \overline{\cos \omega t}, \\ \Lambda_s(f) &= \overline{t \sin \omega t}, & \Lambda_c(f) &= \overline{t \cos \omega t}, \\ \Xi_s(f) &= \overline{t^2 \sin \omega t}, & \Xi_c(f) &= \overline{t^2 \cos \omega t}, \end{aligned} \quad (\text{A1})$$

and with the similar overtone quantities, calculated at the frequencies $2f, 3f, \dots, 2nf$. Now our goal is to show that under certain conditions the quantities $\Omega_{c,s}, \Lambda_{c,s}, \Xi_{c,s}$ have small magnitude in comparison with the quantities $1, \overline{t}, \overline{t^2}$, resp., and thus can be neglected. Let us assume that at the given frequency f the phases ωt_i are distributed approximately uniformly in the segment $[0, 2\pi]$. This means that the multipliers $\cos \omega t$ and $\sin \omega t$ in (A1) may be considered as random quantities. Their values are jumping randomly in the segment $[-1, +1]$, whereas the functions $1, t$, and t^2 are varying slowly. Therefore, the mentioned sines and cosines may be treated as random quantities not correlated with the timings t_i . This quasirandom property allows us to write down approximations like $\overline{\sin \omega t} \sim 1/\sqrt{N}$ and $\overline{t \sin \omega t} \approx (\overline{t})(\overline{\sin \omega t}) \sim \overline{t}/\sqrt{N}$.

Therefore, all the quantities (A1) may be expected to be negligible (at the given frequency f) when the values of $\Omega_{c,s}(f)$ are small. It is not hard to see that $\Omega(f) = \Omega_c(f) + i\Omega_s(f) = e^{i\omega t}$ (with i being the imaginary unit) represents the complex spectral window of the time series and the square of its module is the usual spectral window. Typically, the spectral window contains a strong narrow peak at $f = 0$ and a series of smaller peaks, corresponding to aliasing frequencies. Therefore, in the case when the spectral window does not contain any strong peaks at the frequencies $f, 2f, \dots, 2nf$, we can keep in the matrices $\mathbf{Q}, \mathbf{S}, \mathbf{R}$ only the terms, which do not contain sines or cosines inside the averaging operation. In this approximation, the matrix \mathbf{M} can

be calculated easily. The result is given in (8), and the eigenvalues required are approximated as $\lambda_{2k} \approx \lambda_{2k-1} \approx \pi T_{\text{eff}}^2 k^2$.

It is not hard to check that when our base model $\mu_{\mathcal{H}}$ is not empty but contains a free constant term or a low-order polynomial drift with free coefficients, the same approximation for the matrix \mathbf{M} holds true under similar conditions. In this case, the base model $\mu_{\mathcal{H}}$ appears approximately orthogonal to the signal model μ in the sense that the cross averages $\overline{\varphi_{\mathcal{H}} \otimes \varphi}$ can be neglected in comparison with the respective elements of the matrices $\overline{\varphi_{\mathcal{H}} \otimes \varphi_{\mathcal{H}}}$ and $\overline{\varphi \otimes \varphi}$.

Coupled with the obtained approximate expressions for the eigenvalues λ_k , the eq. (7) yields the eq. (9) with

$$\begin{aligned} \alpha_n &= \frac{1}{2\pi\Gamma(n + \frac{1}{2})} \int_0^\infty \left(1 - \frac{1}{\prod_{k=1}^n (1 + xk^2)}\right) \frac{dx}{x^{3/2}} = \\ &= \frac{2^n}{(2n-1)!!} \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left(1 - \frac{1}{\prod_{k=1}^n (1 + x^2k^2)}\right) \frac{dx}{x^2}. \end{aligned} \quad (\text{A2})$$

We can calculate the integral in (A2) using the theory of functions of a complex variable. Denoting the integrand in the last integral in (A2) as $f(x)$, we can easily check that $\lim_{x \rightarrow \infty} |xf(x)| = 0$ (where x is considered as a complex variable). This means that we can replace the integration line $(-\infty, +\infty)$ by a closed contour \mathcal{C}_R , representing a semi-circle of the radius $R \rightarrow \infty$ in the upper complex semiplane. Indeed, the integral over the semicircle arc decays at least as rapidly as $\sim \pi R f(R) \sim \pi/R \rightarrow 0$ when $R \rightarrow \infty$, and the integral over the diameter of the semi-circle, $(-R, R)$ tends to the integral within $(-\infty, +\infty)$ that we need to compute.

The integrand $f(x)$ can be represented as a ratio of two algebraic polynomials: $f(x) = P(x)/Q(x)$, where $Q(x) = \prod_{k=1}^n (1 + x^2k^2)$ and $P(x) = (Q(x) - 1)/x^2$ (it is not hard to see that $P(x)$ is indeed a polynomial of degree $2n-2$, because the free term in $Q(x)$ is unit and hence the denominator x^2 is reduced). Therefore, the integral over \mathcal{C}_R can be expressed via the sum of residues of $f(x)$ in the points $x_k = i/k$, $k = 1, 2, \dots, n$ (with i being the imaginary unit), which represent the roots of $Q(x)$ in the upper complex semiplane. That is,

$$\frac{1}{2\pi i} \int_{-\infty}^{+\infty} f(x) dx = \frac{1}{2\pi i} \lim_{R \rightarrow \infty} \int_{\mathcal{C}_R} f(x) dx = \sum_{k=1}^n \text{Res} f(x_k). \quad (\text{A3})$$

Since the singularities x_k are simple poles, the corresponding residues can be evaluated as

$$\begin{aligned} \text{Res} f(x_k) &= \frac{P(x_k)}{Q'(x_k)} = \frac{k}{2i \prod_{j=1..n, j \neq k} (1 - j^2/k^2)} = \\ &= \frac{k^{2n-1}}{2i \prod_{j=1..n, j \neq k} (k^2 - j^2)} = \frac{(-1)^{n-k} k^{2n+1}}{i(n+k)!(n-k)!}. \end{aligned} \quad (\text{A4})$$

The formulae (A2,A3,A4) yield the final expression (10).

Note that alternatively we could use the treatment involving ellipsoidal surfaces in multi-dimensional spaces (see eq. (B7) by Baluev 2008). This way seems to be less convenient to obtain exact formulae for α_n , but nonetheless it yields a simple upper bound

$$\alpha_n \leq \frac{1}{(n-1)!} \sqrt{\frac{1}{n} \sum_{k=1}^n k^2} = \frac{1}{(n-1)!} \sqrt{\frac{(n+1)(2n+1)}{6}}. \quad (\text{A5})$$

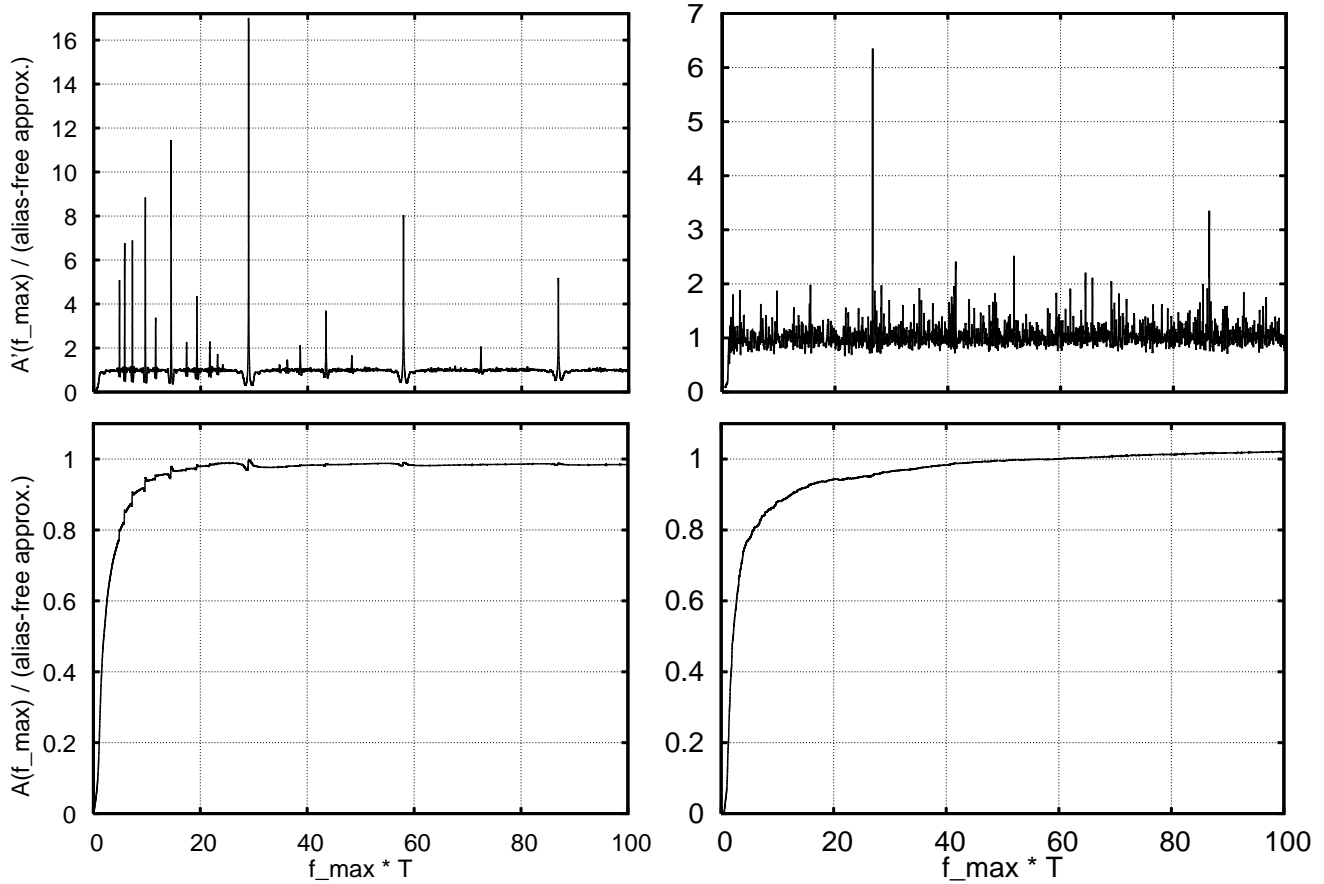


Figure A1. The figure shows the precision of the alias-free approximation of the factor $A(f_{\max})$. Top-left panel: the graph of the ratio of the derivative $A'(f_{\max})$ (the inner integral in (7)) to its alias-free approximation $2\pi^{n+0.5}\alpha_n T_{\text{eff}}$. On an almost horizontal graph, we can see a sequence of strong but narrow splashes corresponding to aliasing periods. Bottom-left panel: the similar graph for the function $A(f_{\max})$ itself. The splashes at the aliasing frequencies exist but are very small and do not produce significant perturbations. The data were obtained for $n = 3$ and $d_{\mathcal{H}} = 1$ (with a free constant term in the model $\mu_{\mathcal{H}}$). The $N = 100$ timings of the mock input time series were periodically gapped with a frequency corresponding to $Tf \approx 28.5$. At this gapping frequency, the folded phases spanned only $\approx 10\%$ of the full period. Right panels show similar graphs for the case of $N = 30$ randomly spaced observations, $n = 5$ and $d_{\mathcal{H}} = 1$.

The comparison of this bound with numerical values from Table 1 shows that this bound is remarkably sharp.

Formally, the approximation (9) was based on certain assumptions of negligible aliasing, which we have discussed above. Nevertheless, it was demonstrated in (Baluev 2008) for the LS periodogram, that this approximation of the factor $A(f_{\max})$ is quite precise in practice, even when the aliasing effects are strong. We may expect the same behaviour of $A(f_{\max})$ for multi-harmonic periodograms. This is due to the integral character of the representation (7). Indeed, the aliasing may result in a strong perturbation of the eigenvalues λ_k and hence of the inner integral in (7). However, these perturbing effects are locked in very narrow frequency intervals of the typical width $\Delta W \sim 1$. After integration over a wide frequency range with $W \gg 1$, the resulting perturbation in the whole integral appear insignificant. This is illustrated in Fig. A1.

Therefore, the only practically important source of a possible inaccuracy of the analytic FAP estimation lies in the possible unsharpness of the Davies bound (4) itself and in the possible inaccuracy of the associated alias-free approximation (5) itself.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.